TASK: "Given RNA sequence features, estimate what type of cancer a patient has"

1. Data Collection

Gene Expression Cancer RNA-seq Data Set was downloaded from "UCI Machine Learning Repository". A total of 801 samples (data points) with 20531 features were found.

Table 1: Sample Composition

Cancer Type	Sample Count
PRAD	136
LUAD	141
BRCA	300
KIRC	146
COAD	78

2. Data Pre-Processing

The data was split into train, test data set using 40% data as test set. As can be seen in Table 1there is imbalance in the data sample for each class label, care was taken that this imbalance is also reflected in the training set.

Expression samples were checked for missing features (i.e. features with no data), no such features were found. The training sample features were normalised by removing the mean and scaling to unit variance.

3. Feature Selection

a. Supervised Feature Selection

All training set features with no variance were removed

All training set features having a pearson correlation coefficient of above 0.70 (positive and negative) i.e. having 70% corelation were filtered out. Using these two methods we reduced the featues from 20531 to 12427

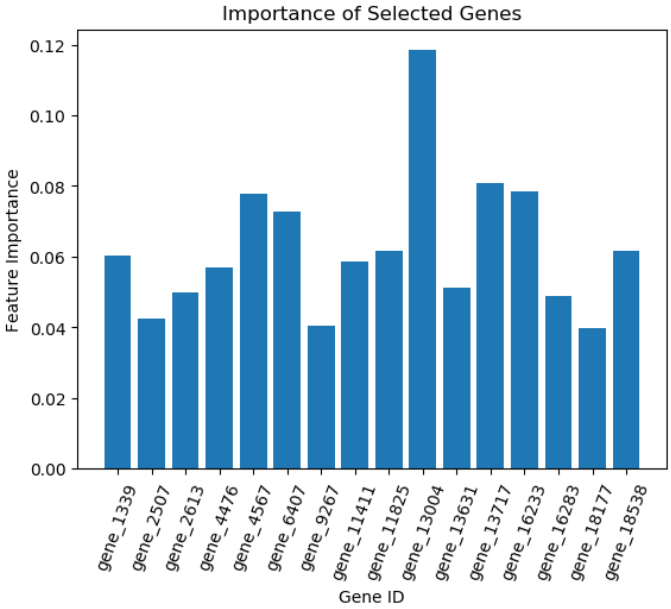

Fig.1: Feature Importance

b. Unsupervised Feature Selection

The training set features were further filtered using multiple iteration of Random Forest Classifier using mean as threshold to select features at each iteration. The iteration were run until a mean cross validation accuracy was found to be lower than 0.98 or the number of

features were below or equal to 15. 16 features were selected. Fig. 1 shows the gene id which were selected and their importance to the model.

4. Classifier Creation

A Support Vector Classifier(SVC) with radial bias was selected for building the classifier based on literature review. For biological data with small sample size multi-class label the performance of SVC is found to be better than other classifiers like Logistic Regression, kNN or LDA/QDA.'

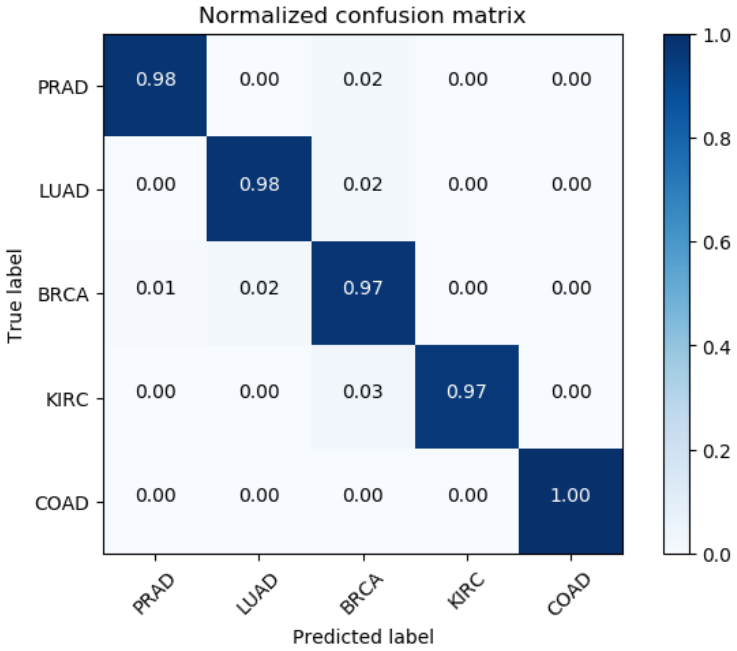

Fig. 2: Normalised Confusion Matrix

5. Classifier Assessment

The classifier is found to have a mean cross-validated training accuracy of 0.99(+/- 0.01) and a mean cross-validated testing accuracy of 0.98(+/- 0.03). Table 2 showcases the classification report for the classifier on test set. A normalised confusion matrix shown in Fig. 2 was also generated to support the overall assessment of the classifier.

Accuracy	0.98			
	Precision	Recall	Score	Support
PRAD	0.98	0.98	0.98	54
LUAD	0.97	0.98	0.97	57
BRCA	0.97	0.97	0.97	120
KIRC	1.00	0.97	0.98	59
COAD	1.00	1.00	1.00	31
Total	0.98	0.98	0.98	321

6. Conclusions

The classifier created shows a very high rate of training and testing accuracy which in theory could be misleading due to the very small nature of the dataset as there are only 801 samples. A more intensive data-set is needed for better training of the model which will automatically increase the reliability of the model.

During the feature selection step a better feature dataset could have been filtered if the gene names were not masked by the gene id, as gene names would have allowed us to do literature review and filter based on that.

7. File Structure

The zip file attached with this email contains the model_maker.py which creates the model and creates the files that are required by the webserver. The model_maker.py takes multiple arguments

$python model_maker.py <correlation_cutoff_value> <input_file_path> <output_file_path>

The folder cancer_predictor contains the django server on which the webapp is hosted.

There is a requirements.txt file which contains the conda environment libraries that are necessary for the running the webapp and model_maker.py